Psychological Test and Assessment Modeling, Volume 56, 2014 (2), 154-167

The influence of different rating scales on impression management in high stakes assessment

Lale Khorramdel[1]

Abstract

The impact of different rating scales on intentional response distortion in personality questionnaires in high stakes assessment was investigated by administering the Personality Research Form (PRF; Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984) to 268 applicants in real selection situations. The applicants responded with either a 6-point rating scale (n = 184) or a 2-point rating scale (n = 84). It was hypothesised that a 6-point rating scale leads to less intentional response distortion than a 2-point rating scale, as it might be more difficult to adjust responses to a faking good schema. Both applicant groups were additionally compared to a volunteer sample (n = 184) randomly selected from the PRF norm sample. Results provide evidence of faking tendencies in the applicant samples and show an advantage of the 6-point rating scale (less faking tendencies). Moreover, it is assumed that the type of response format might interact with item content and wording. Nevertheless, even the applicant group with the 6-point rating scale seems to have faked responses compared to a volunteer sample.

Key words: high stakes assessment, impression management, rating scales, response behaviour

[1] *Correspondence concerning this article may be addressed to:* Lale Khorramdel, PhD, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria; email: lale.khorramdel@univie.ac.at

Introduction

Response formats are not only used to present response alternatives to measure a given construct, but also to moderate item characteristics such as item difficulty or guessing probability in cognitive ability or achievement tests, as well as item difficulty and transparency in personality questionnaires. Adjusting the difficulty and transparency of items in personality questionnaires is especially important, since their high transparency often makes the measured construct evident to the test-taker (Furnham, 1986), which in turn makes the questionnaires prone to faking (Dilchert, Ones, Viswesvaran, & Deller, 2006; Viswesvaran & Ones, 1999; Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007). The use of certain response formats is a promising attempt to reduce faking in personality questionnaires, which are being used increasingly by organizations in order to select the most suitable job applicants (especially if applicants show low variance in their cognitive abilities). The problem of faking behaviour (intentional response distortion, impression management) in this context is obvious and amplified by the fact that it occurs in different faking styles (Robie, Brown, & Beaty, 2007; Zickar, Gibby, & Robie, 2004), varying between applicants along with their ability or perceived ability to fake and their motivation to do so (cf. Goffin & Boyd, 2009; Snell, Sydell, & Lueke, 1999). Hence, rank order changes take place that influence which applicant gets hired (Griffith, Chmielowski, & Yoshita, 2007; Mueller-Hanson, Heggestad, & Thornton, 2003; Robie, Brown, & Beaty, 2007; Winkelspecht, Lewis, & Thomas, 2006). Because personality measures are believed to provide important information in addition to cognitive measures and to improve the validity of the selection process (Kuncel, Hezlett, & Ones, 2001; Kyllonen, Walters, & Kaufman, 2005; Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007a, 2007b; Schmidt & Hunter, 1998), they are often used in high stakes assessments. High stakes assessments are assessments in which the test result leads to important consequences for the individual test-taker (in contrast to low stakes assessment where this is not the case). As they are favoured measures when it comes to assess personality there is an impetus to optimize them by reducing their vulnerability to faking, for instance by examining different response formats.

It is assumed that forced-choice response formats (e.g. being asked to choose between different options of behaviour) minimize faking tendencies in contrast to single-stimulus response formats (rating scales) where the degree of agreement for a statement has to be expressed by marking exactly on a number of (item-non-specific) ordered response categories, which are typically the same for all statements or items measuring the same construct. It is hypothesized that forced-choice response formats are less transparent and make it more difficult to respond desirably (Jackson, Wroblewski, & Ashton, 2000; Martin, Bown, & Hunt, 2002). Nevertheless, it has been shown that test-takers are able to distort their responses using a forced-choice format as well (Lammers & Frankenfeld, 1999) and that a forced-choice response format is not better at retaining the rank ordering of individuals in comparison to a single-stimulus response format (Christiansen, Burns, & Montgomery, 2005; Heggestad, Morrison, Reeve, & McCloy, 2006). Comparing response formats with different numbers of response options has shown that dichotomous response formats (where participants choose one of two alternatives) lead to higher fak-

ing tendencies (more precisely impression management and intentional response distortion) in contrast to analogue scales, in which participants mark the extent of their agreement along a continuous line between two alternatives (Khorramdel & Kubinger, 2006; Kubinger, 2002; Seiwald, 2002). Furthermore, a dichotomous response format might provoke reactance, resulting in atypical or arbitrary responses that do not describe the subjects' true character (Karner, 2002). Hence, both reactance and the ability to fake responses might be reduced by providing a larger number of response options, as is done to reduce guessing effects in ability tests.

Based on these findings, the current study examines the influence of different numbers of response options in rating scales by comparing a 2-point rating scale (1 = false and 2 = true) with a 6-point rating scale (1 = disagree totally to 6 = agree completely). Furthermore, the study has the advantage of testing applicants in real selection situations (high stakes assessments), where the motivation to make a good impression is assumed to be high since the test result leads to consequences important to the applicants. The current study focuses on rating scales because they are the most popular response formats in psychological assessment: they provide the possibility of interindividual comparisons (Heggestad, Morrison, Reeve, & McCloy, 2006), smaller effort in item construction (compared to forced-choice formats), and respondents are able to focus on only one common response format which facilitates the response process.

Hypothesis

According to the findings of Khorramdel and Kubinger (2006), we expect that, similar to an analogue scale, the 6-point rating scale will be less vulnerable to intentional response distortion than a 2-point rating scale. Test-takers' ability to adjust their responses to a faking good schema or to give stereotypical responses (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) should be higher with a 2-point rating scale, while a 6-point rating scale should force test-takers to consider their responses more precisely.

Moreover, both applicant samples are compared to volunteers randomly selected from the PRF norm sample and it is expected that higher faking tendencies are found in the applicant samples.

Using findings of faking studies for the interpretation of current results

In order to find out which scores from the PRF might show faking tendencies (with regard to the different questionnaire scales) the findings of a few faking studies that used the German edition of the PRF were considered. According to a study from Stumpf and Steinhart (1981) who used the German edition of the PRF to investigate the effects of faking-good and faking-bad instructions on a sample of soldiers and officers in training with the German Armed Forces, the following findings can be reflected upon: 1) *faking-good instructions* led to increased scores in the PRF scales Achievement, Affiliation, Dominance, Endurance, Exhibition, Nuturance, Order, Social Recognition, Succorance,

and Understanding, and to decreased scores in the PRF scales Aggression, Harm avoidance, Impulsivity, and Play, in contrast to faking-bad instructions or standard instructions; 2) *faking-bad instructions* led to decreased scores in the PRF scales Achievement, Affiliation, Dominance, Endurance, Exhibition, Nurturance, Order, Social Recognition, Succorance, and Understanding, as well as to increased scores in the PRF scales Aggression, Harm avoidance, Impulsivity, and Play, in contrast to faking-good instructions or standard instructions. Altogether, faking-bad instructions led to higher differences in the scores than faking-good instructions. Rather similar effects were found in studies that used the English version of the PRF (Braun & Asta, 1969; Braun & Constantini, 1970; Hoffmann, 1968; Hoffmann & Nelson, 1971; Holden & Jackson, 1981) except for the scale Harm avoidance, where contrary results were found with respect to faking bad-instructions.

It may be true that applicants do not fake as much as volunteers under faking instructions (Stumpf & Steinhart, 1981; Viswesvaran & Ones, 1999) but the direction of how dimensions are faked should be quite the same. As the sample in the faking study of Stumpf and Steinhart (1981) were German speaking soldiers like the applicants in the current study, it should be possible to use the results of Stumpf and Steinhart (1981) to examine which applicant group was faking which scale to a higher extent.

Method

A study within personnel selection is presented to investigate if rating scales can be optimized by applying more response options (6-point rating scale) in contrast to only two response options (2-point rating scale).

Measures

Personality Research Form (PRF) – German edition. The PRF (Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984) is a personality questionnaire based on Murray's personality theory (1938) that measures a set of traits important for psychological research as well as psychological assessment. 234 items with a dichotomous response format ("true" and "false") measure fifteen scales: Achievement, Affiliation, Aggression, Dominance, Endurance, Exhibition, Harm Avoidance, Impulsivity, Nurturance, Order, Play, Social Recognition, Succorance, Understanding, and Infrequency. The Cronbachs alpha reliabilities of the PRF scales, presented in the test-manual, range from .69 to .87.

Sample and design

The 268 applicants were all soldiers of the Austrian Federal Armed Forces who had applied for pilot training. The sample consists almost only of men (98.8%) between the age of 18 and 23 (mean age: 19.67 years; SD = 1.134) with German as their first lan-

guage and education levels varying from a compulsory education of 9 years (13%) and apprenticeship (43%) to general qualification for university entrance (44%). Furthermore, all applicants had undergone a pre-selection process (physical and psychological examination, basic cognitive tests measuring constructs such as concentration, perception speed, reasoning…) in the Austrian Federal Armed Forces, before they applied for the training. 184 applicants received the PRF as computer based assessment with a 2-point rating scale while 84 applicants received the PRF as paper based assessment with a 6-point rating scale; the data were collected in 2008 and 2009. (The selection process worked with different cycles; applicants were compared to each other within single cycles. To not influence the rank order of applicants within a cycle by applying different response formats, they could not be randomly assigned to the two different groups but were tested in subsequent cycles.)

To test if the applicants in the current study responded differently compared to volunteers (control group), the data from the current study were compared to data of the PRF norm sample (provided by the authors of the PRF). For this purpose 184 male respondents with a mean age of 30.82 (SD = 11.26; 31.5% between 18 and 23 years, 27.7% between 24 and 30 years, 26.1% between 31 and 45 years) were selected from the PRF norm sample by using the "select random sample of cases" provided in SPSS (no information about the educational level of the volunteer sample could be obtained). In a first step the PRF norm sample was split into two subsamples by gender, and then the 184 cases were randomly selected from the male subsample in order to provide a comparable volunteer sample to the applicant sample with regard to gender.

Data collection and data preparation

First, all applicants attended a psychological assessment carried out by the department of Human Resources of the Austrian Federal Armed Forces, where their cognitive abilities and personality traits were tested. After an assessment of approximately 2 x 4 hours (4 hours testing, 2 hours break, again 4 hours testing) of working on cognitive ability and achievement tests, they filled out either the paper based version of the PRF by responding to a 6-point rating scale, or the computer based version of the PRF with a 2-point rating scale. None of the applicants had received any information about the requirements profile.

To be able to compare the two applicant groups in our analysis, the 6-point rating scale was scored dichotomously (post-hoc dichotomisation; categories 0, 1, and 2 were recoded to 0, and categories 3, 4, and 5 were recoded to 1), so that marks on one side indicated only agreement or disagreement. The sum score for each questionnaire scale was used in the analysis.

Results

To investigate the effects of the two different rating scales by comparing the means of the two applicant groups with one another a multivariate analysis of variance (MANOVA) was conducted with the main factor Response Format considering $\alpha = .05$ for interpretation of results. With $\alpha = .05$ and $\beta = .20$ an ANOVA is able to detect a mean difference of $\delta \geq 2/3\ \sigma$ (the standard deviation of each scale) by testing 37x2 = 74 subjects (as a matter of fact we had to calculate the sample size for ANOVA). With either 84 or 184 test-takers per group adequate sample sizes had been realised. In addition, both applicant groups were compared to a random subsample of the PRF norm sample (provided by the PRF authors) to investigate if the applicants' response behaviour differs significantly from those of volunteers by conducting Welch-Tests (requirements for computing a MANOVA were not met). To counteract the problem of multiple comparisons within one sample a Bonferroni adjustment was used to control for familywise error rate.

Results of the MANOVA for the main factor Response Format comparing applicants

The means and standard deviations of all scales in each group are given in Table 1. Box's M-Test for testing the homogeneity of the variance-covariance matrix was significant ($p = .003$). To ascertain whether this significance is due to particular dependent variables (questionnaire scales) on account of the heterogeneous variances, Levene's test was calculated for each scale. The five scales Achievement, Aggression, Order, Succorance, and Infrequency were disclosed to be significant in Levene's test ($p = .009$, $p = .000$, $p = .040$, $p = .044$, $p = .022$). After excluding these scales Box's M-Test proved to be non-significant ($p = .276$). That is, the resulting F-values of the multivariate analysis of variance can be fairly interpreted. The MANOVA for testing the main effect of the factor Response Format shows a significant effect of the Response Format ($p < .001$; $F = 6.704$; hypothesis $df = 10$; error $df = 257$; $\eta^2 = .207$). Considering $\alpha = .05$, univariate factorial ANOVAs of each single scale show significantly different means of the five scales Affiliation ($p = .021$), Endurance ($p = .010$), Harm Avoidance ($p = .017$), Social Recognition ($p < .001$), and Understanding ($p < .001$) between the two applicant groups. Considering a Bonferroni adjustment for the ten scales comprised in the MANOVA (.05 / 10 = .005), only the scales Social Recognition and Understanding show to be significant with p-values smaller than .005. See in Table 1 the respective means. To additionally investigate the effect of the factor Response Format on the scales Achievement, Aggression, Order, Succorance, and Infrequency, which had to be excluded from the MANOVA, Welch-Tests were applied. Computing a Bonferroni adjustment for these 5 scales (.05 / 5 = .01), p-values smaller than .01 are considered to show significant mean differences. While significant effects resulted for the scales Aggression ($p = .008$), Order ($p = .008$), and Succorance ($p = .002$), no significant effect occurred with regard to the scales Achievement ($p = .118$) and Infrequency ($p = .027$).

Table 1:
Means and standard deviations for all dependent variables (scales) with regard to the factor Response Format; applicant groups and volunteers

Dependent Variable	*2-Point Rating scale* (n=184)		*6-Point Rating scale* (n=84)		*Volunteers* (n=184)										
	group-1		*group-2*		*group-3*		*group-1 versus group-2*			*group-1 versus group-3*			*group-2 versus group-3*		
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	*F* or *T*	*p*	η^2	*T*	*p*	η^2	*T*	*p*	η^2
Achievement	13.52	1.41	13.15	1.92	10.33	3.08	1.57 (T)	.118	.019	12.74	.000	.387	9.04	.000	.256
Affiliation	14.35	2.18	13.67	2.30	9.18	3.87	5.43	.021	.020	15.81	.000	.463	11.76	.000	.361
Aggression	4.32	2.13	5.24	2.78	5.43	2.90	-2.68 (T)	.008	.053	-4.13	.000	.048	-.63	.529	.002
Dominance	12.34	2.83	13.04	2.75	7.59	3.78	3.52	.062	.013	13.64	.000	.354	13.20	.000	.451
Endurance	13.67	2.20	12.89	2.45	8.78	3.52	6.66	.010	.024	16.10	.000	.459	10.97	.000	.354
Exhibition	10.16	2.99	10.77	2.64	7.59	3.62	2.58	.110	.010	7.46	.000	.136	7.98	.000	.231
Harm Avoidance	3.75	2.63	2.91	2.68	7.46	3.95	5.81	.017	.021	-10.58	.000	.260	-10.97	.000	.005
Impulsivity	4.91	2.61	4.73	2.25	7.70	3.26	.32	.571	.001	-9.04	.000	.189	-8.67	.000	.254
Nurturance	11.72	2.31	11.65	2.43	9.21	3.23	.05	.824	.000	8.57	.000	.181	6.87	.000	.186
Order	12.15	3.16	13.16	2.72	8.31	4.29	-2.67 (T)	.008	.037	9.87	.000	.225	11.19	.000	.347
Play	9.76	3.12	9.32	2.86	8.39	3.52	1.21	.273	.005	3.92	.000	.041	2.40	.018	.029
Social Recognition	8.16	2.95	9.63	2.61	7.94	3.84	15.33	.000	.054	.52	.601	.000	4.10	.000	.069
Succorance	7.28	2.58	6.31	2.27	6.25	3.35	3.12 (T)	.002	.051	3.27	.000	.030	.25	.801	.001
Understanding	9.84	2.72	8.41	3.06	9.80	3.19	14.64	.000	.052	.12	.904	.000	-3.31	.000	.063
Infrequency	.46	.66	.69	.85	–	–	-2.24 (T)	.027	.037	–	–	–	–	–	–

Results of the Welch-Tests comparing applicants and volunteers

As a MANOVA could not be conducted to compare volunteers with applicants because the requirement of homogeneity of the variance-covariance matrix was not met (analysis showed a significant Box's M-Test with $p < .001$, and almost all questionnaire scales showed to be significant in Levene's test with $p \leq .001$ to $p = .021$, except for three scales with $p = .064$ to $p = .330$), multiple Welch-Tests were computed. The scale Infrequency was not included in the analysis as its items differ between the PRF version used in the norm study and the PRF version used in the current study leaving 14 scales (instead of 15) in the analysis. Computing a Bonferroni adjustment for $\alpha = .05$ for the 14 scales as dependent variables ($.05 / 14 = .0035$), p-values smaller than .0035 are considered to show significant mean differences. Results for the applicant group with the 2-point rating scale compared to the volunteer group show significant differences in almost all scales (with $p \leq .001$) except for the scales Social Recognition ($p = .601$), and Understanding ($p = .904$). Comparing the applicant group with the 6-point rating scale with the volunteer group results show significant differences in almost all scales (with $p \leq .001$) except for the scales Aggression ($p = .529$), Play ($p = .018$), and Succorance ($p = .801$). Means and standard deviations for all groups are listed in Table 1.

Interpretation

The results of the MANOVA reveal a significant main effect of the factor Response format on an applicant's response behaviour. The independent analyses of the single scales show that five scales of the PRF are affected by the factor Response format. The means of the applicant groups in each scale (see Table 1) are interpreted according to Stumpf and Steinhart's findings (1981) in order to ascertain which response format may lead to fewer faking tendencies. Hence, higher scores in the scales Order, Social Recognition, Succorance, and Understanding, as well as lower scores in the scales Aggression may show faking tendencies in the sense of response distortion when the two applicant groups are compared.

Applicants who had been given the 2-point rating scale showed higher scores in the scales Succorance, and Understanding, as well as lower scores in the scale Aggression than subjects who had been given the 6-point rating scale. These results lead to the assumption that the 2-point rating scale provokes higher faking tendencies than the 6-point rating scale which seems to lead to fewer faking tendencies. The opposite is the case with the scales Order, and Social Recognition, where it seems that a 2-point rating scale led to less faking tendencies than the 6-point rating scale. Applicants with a 6-point rating scale showed higher scores in the scales Order, and Social Recognition. Our interpretation is illustrated in Table 2.

Table 2:
Interpretation of the means of the five significant dependent variables (scales) with regard to the factor Response Format (two applicant groups) according to Stumpf and Steinhart (1981), and Hoffmann (1968); applicant groups

Dependent Variable (Scale)	*2-Point Rating Scale* (n=184)	*6-Point Rating Scale* (n=84)
	Mean	Mean
Aggression	4.321	5.238
Succorance	Faking 7.283	6.310
Understanding	9.837	8.411
Order	12.147	Faking**?** 13.155
Social Recognition	8.163	9.631

Comparing both applicant groups with a group of volunteers (coming from the PRF norm sample), results are showing that the two applicant groups show higher faking tendencies in almost all scales according to Stumpf and Steinhart (1981). Both applicant groups show higher scores in the scales Achievement, Affiliation, Dominance, Endurance, Exhibition, Nurturance, and Order, as well as lower scores in the scales Harm Avoidance, and Impulsivity (whereas the scales Social Recognition showed significant differences only for applicants with the 6-point rating scale, and the scales Aggression and Succorance only for applicants with the 2-point rating scale; see Table 1).

Discussion

A study was conducted to investigate the influence of two different rating scales on intentional response distortion within a personnel selection situation (pilot applicants in the Federal Armed Forces) in order to examine if rating scales can be optimized by using a higher number of response alternatives: a 6-point rating scale (with 1 = disagree totally to 6 = agree completely) was compared to a 2-point rating scale (with 1 = false and 2 = true). The MANOVA shows a significant main effect; the factor Response Format affected applicants' response behaviour in nine scales of the PRF. According to the means in the scales Succorance, Understanding, and Aggression, applicants showed less intentional response distortion or faking tendencies when they responded with a 6-point rating scale than when responding with a 2-point rating scale. These findings resemble those of Khorramdel and Kubinger (2006), who were able to show that (normative) items were less fakable when answered on an analogue scale than with a dichotomous response format. Similar to forced-choice formats (Jackson, Wroblewski, & Ashton, 2000; Martin, Bown, & Hunt, 2002) rating scales with more response options seem to make it more difficult to fake responses or to adjust responses to a faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992).

The means of the scales Order, and Social Recognition showed the opposite effect: a 2-point rating scale seems to show less faking tendencies than the 6-point rating scale. Upon closer inspection of the content of the items of these two scales, it seems that these items were more transparent than items from the other PRF scales: the scale Order includes many items concerning challenges and tasks that are considered daily routines in the Federal Armed Forces or for pilots (e.g. making plans, hanging up clothes, arranging things, attaching importance to one's appearance); a similar explanation can be applied to the scale Social Recognition (e.g. the importance of prestige, image or reputation, as well as acceptance). Since items with transparent content or content that is of particular importance to a specific job have been found to be more fakable (Furnham, 1986; Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006), we assume that the items of these two scales were highly vulnerable to faking tendencies, and that not even the 6-point rating scale could help to make the desirable response less transparent. Of course, this only explains why the 6-point rating scale was not less fakable, not why the 2-point rating did not fulfil our expectations. What is noticeable, apart from the item content regarding the scales Order, and Social Recognition, is that many items of these scales contain extreme phrases like "never", "always" or "almost always", "by no means", or "inexcusable". In combination with the transparent item content, these extreme phrases might have led to some kind of reactance when applicants had to respond with a 2-point rating scale, in the sense that some statements that might have been affirmed when presented with more moderate answer possibilities (that is for example provided with the 6-point scale) were instead refused. In this respect, we assume that the decreased socially or job-related desirable responses are not to be interpreted as decreased fakability of items with a 2-point rating scale but rather as an underestimation of the actual trait loading.

However, comparing the two applicant samples to a volunteer sample (a random sample drawn from the PRF norm sample) showed significant mean differences for both rating scales in most PRF scales which may be interpreted as faking tendencies. Thus, not only the 2-point rating scale led to faking tendencies, but also the 6-point rating scale.

In summary, the 6-point rating scale seems to be a better solution than the 2-point rating scale, as less faking tendencies were revealed in most of the PRF scales; it might be less evident for the test-taker what is socially or job-related desirable when more response options are given. But this might not be true for all scales, as this effect seems to be bound to the scale or item content. We assume that response format interacts with item content and that items should be developed or used (with regard to their content) very carefully. We also assume that a 2-point rating scale not only enhances faking tendencies, but might also harm the measurement. Nevertheless, the fact that the 6-point scale showed less intentional response distortion in most scales does not mean that no intentional response distortion occurred. Results from comparing the two applicant samples with a volunteer sample lead to the assumption that responses were also faked in the condition with the 6-point rating scale. Thus, using rating scales with more than two response options might provide more advantages but does still not solve the problem of intentional response distortion in selection situations, nor the problem of different faking styles which are assumed to change the rank order of applicants. Further studies investigating possible interactions between different response formats and faking styles would

be interesting. However, our study was able to show once again that the response format has a moderating effect with respect to intentional response distortion.

Limitations and implications for further research

While we assume that the combination of the job-specific, transparent item content with the extreme phrases might have led to reactant response behaviour when the 2-point rating scale was applied, this remains only an assumption. Further research may examine this assumption more closely and find other or better explanations. Further research might also pay more attention to the effects of the combination of different questionnaire administration modes on response behaviour, as well as the interaction of such variables with the item content and item wording.

As all applicants were men who had applied for a pilot training and who all came from the same institution (Austrian Federal Armed Forces), the findings might not apply to other occupational groups or women. Future research should investigate whether our results can also be found in other samples.

The fact that the PRF was administered differently to the two applicant groups (paper based versus computer based assessment) can be criticised as well but prior studies could show that there is no notable difference between these assessment modes when it comes to personality questionnaires and faking (c.f. Bader, Hofmann, & Kubinger, 1993). Additionally, it has to be pointed out that the volunteer sample consisted of a broader age range than the applicant samples, whereas one third in the volunteer sample was of the same age, and no information about their educational levels could be obtained. Therefore, the comparison between the applicant and volunteer sample has to be done carefully. Despite these critical points, a major advantage of our study is that we were able to study response behaviour of applicants in real selection situations, which allows a valid estimation of the faking problems' true dimension.

Acknowledgments

We are grateful to the Department of Human Resources of the Austrian Federal Armed Forces, in particular Michael Mikas (Head of the Ambulance for Aviation & Traffic Psychology), Christian Czihak (Head of the Department for Aviation & Traffic Psychology), and Christian Langer (Head of the Department of Human Resources), for granting permission to test applicants within the local selection procedure. Our sincere thanks go to Dr. Fritz Ostendorf and Prof. Dr. Alois Angleitner, of the University of Bielefeld who provided us with (a large part of) the PRF norm sample. We also thank Alexander Uitz for his assistance with data collection.

References

Bader, P., Hofmann, K., & Kubinger, K. D. (1993). Zur Brauchbarkeit der Normen von Papier-Bleistift-Tests für die Computervorgabe: Ein Experiment am Beispiel des Gießen-Tests [On the usefulness of norms of paper-pencil tests for computer-based administration: An experiment using the Gießen-Test]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 129-135.

Braun, J. R., & Asta, P. (1969). Changes in Personality Research Form scores (PRF, Form A) produced by faking instructions. *Journal of Clinical Psychology, 25*, 429-430.

Braun, J. R., & Constantini, A. (1970). Faking and faking detection on the Personality Research Form, AA. *Journal of Clinical Psychology, 26,* 516-518.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item format for applicant personality assessment. *Human Performance, 18*, 267-307.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self-assessments? *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 209-225.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385-400.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151-160.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*, 341-355.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

Hoffmann, H. (1968). Performance on the Personality Research Form under desirable and undesirable instructions: Personality disorders. *Psychological Reports, 23*, 507-510.

Hoffmann, H., & Nelson, J. I. (1971). Desirability responses in the Personality Research Form by a sample of alcoholics. *Psychological Reports, 29*, 559-562.

Holden, R., & Hibbs, N. (1995). Increment validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality, 29,* 362-372.

Holden, R. R., & Jackson, D. N. (1981). Subtlety, information, and faking effects in personality assessments. *Journal of Clinical Psychology, 37*, 379-386.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63,* 272-279.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.

Karner, T, (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*], *44,* 42-49.

Khorramdel. L., & Kubinger, K. D. (2006). The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48,* 378-397.

Kubinger, K.D. (2002). On faking personality inventories. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44,* 10-16.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 27*, 162-181.

Kyllonen, P. C., Walters, A., & Kaufman, J. (2005). Noncognitive constructs and their assessment in graduate education: a review. *Educational Assessment, 10*, 153-184.

Lammers, F., & Frankenfeld, V. (1999). Effekte gezielter Antwortstrategien bei einem Persönlichkeitsfragebogen mit "forced-choice"-Format [Effects of selective response strategies in a personality questionnaire with forced-choice format.] *Diagnostica, 45,* 65-68.

Martin, B. A., Bown, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32,* 247-256.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029-1049.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. III (2006). Individual differences in impression management: An exploratory of the psychological process underlying faking. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 288-312.

Ramsay, L. J., Schmitt, N., Oswald, F. L., Kim, B. H., & Gillespie, M. A. (2006). The impact of situational context variables on responses to biodata and situational judgement inventory items. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 268-287.

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489-509.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], 44, 17-23.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N., & Beloch-till, H. (1984). *Deutsche Personality Research Form (PRF)* [*German Version of the Personality Research Form (PRF)*]. Manual, Göttingen: Hogrefe.

Stumpf, H., & Steinhart, I. (1981). *Zur Anfälligkeit der Skalenwerte der deutschen "Personality Research Form" (KA) gegenüber tendenziöser Beantwortung* [*On the susceptibility of the scales in the German "Personality Research Form" (KA) towards tendentious response behaviour*]. Wehrpsychologische Untersuchungen, Heft 3.

Viswesvaran, C., & Ones, D.S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*, 168-190.

Ziegler, M., Schmidt-Atzert, L., Bühner, M. & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: questionnaire, semi-projective, and objective. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 49*, 291-307.